Generative Probabilistic Model for Detecting Selection on Dispersed Genomic Elements from Polymorphism and Divergence

Ilan Gronau[1], Leonardo Arbiza[1], and Adam Siepel[1]

[1]Department of Biological Statistics and Computational Biology,
Cornell University, Ithaca, NY 14853, USA

Submission type: Research Article

Keywords:

Running Head: Inferring Selection in Dispersed Genomic Elements

Corresponding Author: Adam Siepel
102E Weill Hall, Cornell University
Ithaca, NY 14853
Phone: +1-607-254-1157
Fax: +1-607-255-4698
Email: acs4@cornell.edu

Abstract

We present a new probabilistic method for measuring the influence of natural selection on a collection of short elements scattered across a genome based on observed patterns of polymorphism and divergence. This is a challenging task for various reasons, including variation across loci in mutation rates and genealogical backgrounds, and the influence of demography on patterns of polymorphism. In addition, accounting for the combined effects of different modes of selection is known to be a serious challenge for tests of selection that use patterns of polymorphism and divergence. Our method addresses these challenges by contrasting patterns of polymorphism and divergence in the elements of interest with those in flanking neutral sites. While this general approach is common to several existing tests of selection, our method improves substantially on these methods by making use of a full generative probabilistic model, directly accommodating weak negative selection, allowing information from many short elements to be combined in a statistically rigorous manner, and integrating phylogenetic information from multiple outgroup species with genome-wide population genetic data. Our model is able to account for of weak negative, strong negative, and strong positive selection, by making a small set of simple assumptions on their separate effects on polymorphism and divergence. We implemented an expectation maximization algorithm for inference under this model and applied it to simulated and real data. Using simulations, we show that our inference procedure effectively disentangles the different modes of selection and provides accurate estimates of the parameters of interest that are robust to demography. We demonstrate an application of our methods to real data by analyzing several collections of human transcription factor binding sites identified using recently generated genome-wide chromatin immunoprecipitation and sequencing data.

Introduction

Evolutionary modeling has become an essential tool in genomic analysis. It is particularly important in the analysis of noncoding regions in large, complex eukaryotic regions, which are rich in regulatory elements but are difficult to examine experimentally, and therefore remain sparsely annotated and poorly understood. Among other things, evolutionary models can be used to measure the fractions of nucleotides in a genome that are likely to have fitness-influencing functions (Mouse Genome Sequencing Consortium, 2002; Chiaromonte et al., 2003; Lunter, Ponting and Hein, 2006), to distinguish functional from nonfunctional sequences (Kellis et al., 2003; Guigó et al., 2003; Siepel et al., 2007), and to detect sequences likely to be responsible for phenotypic differences between species (Pollard et al., 2006; Prabhakar et al., 2008). More generally, they provide valuable information about the relative roles of neutral drift, positive selection, and negative selection in determining rates and patterns of sequence evolution.

Most evolutionary analyses of noncoding elements so far have made use of sequence conservation between genomes that diverged millions of years ago. However, many confounding effects limit the utility of these approaches, including turnover of regulatory elements (Dermitzakis and Clark, 2002; Moses et al., 2006; Schmidt et al., 2010), challenges in orthology identification, and alignment error. In principle, data describing genetic variation could help to address these limitations, because it reflects evolutionary processes on much shorter timescales, during which turnover should be much less prevalent. Orthology identification and alignment are also much more straightforward on these time scales. It is well known that patterns of polymorphism within a species and divergence between species can be used to tease apart the effects of positive selection, negative selection, and neutral drift for a given collection of functional elements (McDonald and Kreitman, 1991; Sawyer and Hartl, 1992; Bustamante et al., 2005). In practice, however, it is technically challenging to extract useful information about noncoding elements from patterns of polymorphism and divergence for various reasons. Many noncoding elements of interest, such as transcription factor binding sites, are quite short (typically $<$10 bp) and polymorphism data tends to be quite sparse, so that most elements contain no informative sites. Furthermore, factors such as variation across loci in mutation rates and time to most recent common ancestry, and the influence of demography on patterns of polymorphism, make it difficult to interpret patterns of polymorphism and divergence in regulatory elements, and prohibit straightforward pooling of data from multiple elements across the genome.

Here we describe a new computational method, called Inference of Natural Selection from Interspersed Genomically coHerent elemenTs (INSIGHT), that is designed to address these challenges. INSIGHT uses the general strategy of contrasting patterns of polymorphism and divergence (P&D) in a collection of elements of interest with those in flanking neutral regions, thereby mitigating biases from demography, variation in mutation rates, and differences in genealogical backgrounds. In this way, it resembles McDonald-Kreitman-based methods for identifying departures from neutrality (McDonald and Kreitman, 1991; Andolfatto, 2005; Sawyer and Hartl, 1992; Smith and Eyre-Walker, 2002). Unlike these methods, however, INSIGHT is based on a generative probabilistic model, accommodates weak negative selection (Charlesworth and Eyre-Walker, 2008), and allows weak information from many short elements across the genome to be pooled efficiently, in a manner that avoids statistical pitfalls arising from pooling counts of site classes (Stoletzki and Eyre-Walker, 2011). Our modelling approach directly addresses variable mutation rates and times to most recent common ancestry along the genome and fully integrates phylogenetic information from multiple outgroup species with genome-wide population genetic data. In other recent work, we have applied INSIGHT in a large-scale analysis of transcription factor binding sites in the human genome, using chromatin immunoprecipitation and sequencing (ChIP-seq) data for 78 human transcription factors (TFs) from the ENCODE project (Myers et al., 2011) and 54 unrelated complete human genome sequences from Complete Genomics (http://www.completegenomics.com/public-data/69-Genomes/). Our focus in this paper is to detail the probabilistic model and inference strategy underlying the method, and examine its performance on simulated data under a range of scenarios. In addition, we provide a more detailed analysis of four of the TFs surveyed in our other work.

Methods

General Approach

We consider a collection of typically short functional elements scattered across the genome. This collection could be defined in many possible ways. For example, it might consist of all binding sites for a certain transcription factor, or all binding sites within a certain distance of a coding gene that belongs to a pathway of interest. Alternatively, it could be defined by sites clustered based on histone modifications or other genome-wide functional assays. We will be interested in inferring the selective forces that have shaped patterns of polymorphism and divergence across this set of elements in a certain population (or species) of interest, which we refer to as the *target population*, during the time since its divergence from other closely related species (Fig. 1A). Polymorphism in the target population is assayed by sampling a collection of complete genomes from individuals of that population. We assume sufficient sequencing depth that genotype frequency can be estimated with high accuracy, but it would not be difficult to extend the model to use statistically inferred genotype frequencies based on low-coverage sequence data (Yi et al., 2010). Divergence is considered with respect to the ancestral genome representing the root of the lineage leading to the target population. The sequence at the ancestral genome is probabilistically determined using a phylogenetic model and the complete genome sequences of several outgroup species.

Rates of polymorphism and divergence at any given element are directly contrasted against the rates at putative neutral sites in regions flanking that element (Fig. 1B). This allows our model to account for variation in mutation rates, as well as changes in the genealogical background of the individuals sampled from the target population. In particular, we assume that sites within a genomic block of a certain size, say 10,000 bp, have the same neutral rates of polymorphism and divergence, and that the rates of polymorphism and divergence in elements of interest within a given block are scaled versions of this neutral rates.

We assume each site can evolve under one of four possible modes of selection: neutral drift (neut), strong negative selection (SN), weak negative selection (WN), or strong positive selection

(SP). Strong selection is expected to dramatically reduce the time it takes for a mutation to reach fixation (for SP sites) or be eliminated (for SN sites), implying that SP and SN sites do not contribute to polymorphism. Weak negative selection, on the other hand, does allow polymorphisms to be maintained for a longer period of time, however, mutations at WN sites are assumed to be maintained at low frequencies and never reach fixation. These basic assumptions imply the following observations:

- SN sites not contribute to polymorphism or divergence.
- SP sites contribute to divergence but not to polymorphism.
- WN sites contribute to low-frequency polymorphism but not to divergence.

These assumptions allow the contributions of different modes of selection to be distinguished to a degree, but they provide only weak information about the selective mode at each site. In particular, the vast majority of sites are expected to be monomorphic and nondivergent, and these sites will typically have fairly high probabilities under any of the four selective modes. In fact, only high-frequency polymorphisms can be unambiguously assigned to a certain category (neut), and these are typically rare. Hence, instead of directly modeling all four selective modes, we group WN, SN, and SP sites into a single class of selected sites. Information regarding the different modes of selection, and in particular WN and SP, will be captured by the relative divergence and polymorphism rates in selected sites compared to the rates at neutral flanking sites.

A central feature of our approach is the way it accounts for weak negative selection, using a simple distinction between polymorphic sites with low derived allele frequency (DAF) and those with high DAF. According to our assumptions, polymorphisms in selected sites are restricted to the low-frequency class, whereas polymorphisms in neutral sites are assumed to be distributed between the two classes according to some discrete categorical distribution. Since the identity of the derived allele (among the observed alleles) is not assumed to be known, our model probabilistically determines the frequency class of a polymorphic site by integrating over possible values of the derived allele. This is done using the probabilistic ancestral genome sequence mentioned earlier (Fig. 1). Note that the distinction between low and high frequencies is necessarily somewhat arbitrary. However, using simulations and real data analysis we are able to show that our inference method is not sensitive to the fluctuations in the definition of this threshold (see **Results**). The coarse-grained representation of the full site frequency spectrum using two categories, 'low' and 'high', should, in principle, make our method less sensitive to the effects of recent demographic changes in the target population. This is another topic we focus on in our simulation study (see **Results**).

Probabilistic Model

The model is hierarchical, with a collection of global parameters, ρ, η, γ, and $\boldsymbol{\beta}$, and a collection of local, or block-specific, parameters, $\boldsymbol{\theta}$, $\boldsymbol{\lambda}$, and $\boldsymbol{\lambda}^{\mathbf{O}}$ (see Table 1). Sites are assumed to be partitioned into a series of genomic blocks, B, and each genomic block, $b \in B$, is associated with three block-specific neutral rate parameters, θ_b, λ_b, and λ_b^O. The full data across all sites is denoted by $(\mathbf{X}, \mathbf{O})$, where the data in each site i consists of the outgroup sequence data, O_i, representing the column in the multiple-sequence alignment of outgroup genomes, and the population sequence data, $X_i = (X_i^{maj}, X_i^{min}, Y_i)$, representing the major and minor alleles observed in the target population at site i, and the frequency category of the minor allele. We assume no more than two alleles are observed in the target population at each site (if only one allele is observed then $X_i^{min} =$ NA), and filter other sites from the analysis. The frequency class of a given site, Y_i, is determined by its *minor* allele frequency (MAF) at that site and a fixed frequency threshold, $f < \frac{1}{2}$, for distinguishing between low and high frequencies: $Y_i =$ M for monomorphic sites (MAF=0), $Y_i =$ L for polymorphic sites with MAF$< f$, and $Y_i =$ H for polymorphic sites with MAF$\geq f$. Note that since $f < \frac{1}{2}$, sites with $Y_i =$ H necessarily have a high derived allele frequency, but for sites with $Y_i =$ L, our model will have to probabilistically determine the frequency class of the derived allele (high or low).

Our model assumes independence of the different genomic blocks and conditional independence of nucleotides within blocks given the block-specific parameters. We distinguish between sites within the elements of interest (*element sites*, for short), E, and putative neutral sites flanking these elements (*flanking sites*, for short), F. Each site, $i \in E \cup F$, is associated with a selection class, $S_i \in \{\text{sel}, \text{neut}\}$ and two hidden ancestral states, A_i and Z_i (Table 2). The graphical model shown in Fig. 2 applies to both element sites and flanking sites, however, S_i is fixed at the "neut" value in flanking sites. The likelihood function can be written as follows:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\zeta}\,;\mathbf{X},\mathbf{O}) \;&\equiv\; P(\mathbf{X}\mid\mathbf{O},\boldsymbol{\zeta}) \;=\\ &\prod_{b\in B}\left[\prod_{i\in F_b}\sum_z\sum_a P(X_i, Z_i=z, A_i=a\mid S_i=\text{neut}, O_i,\boldsymbol{\zeta})\right]\\ &\qquad\times\left[\prod_{i\in E_b}\sum_{s\in\{\text{neut,sel}\}} P(S_i=s\mid\boldsymbol{\zeta})\sum_z\sum_a P(X_i, Z_i=z, A_i=a\mid S_i=s, O_i,\boldsymbol{\zeta})\right]\;, \qquad (1)\end{aligned}$$

where F_b and E_b denote the element and flanking sites in block b, respectively. The probability of selection in element sites is determined by a two-component mixture model, with coefficient ρ. The parameter ρ serves a key role in our model, describing the fraction of sites under selection within the functional elements.

$$P(S_i = s\mid\boldsymbol{\zeta}) = \begin{cases}\rho & s = \text{sel}\\[1em] 1-\rho & s = \text{neut}\end{cases}\;. \qquad (2)$$

According to the conditional independence assumptions of the graphical model, each term of the form $P(X_i, Z_i, A_i \mid S_i, O_i, \boldsymbol{\zeta})$ in Equation (1) can be expressed as the following product of conditional probabilities:

$$P(X_i,Z_i,A_i\mid S_i,O_i,\boldsymbol{\zeta}) \;=\; P(Z_i\mid O_i,\;\lambda_b^O)\;P(A_i\mid S_i,Z_i,\;\boldsymbol{\zeta})\;P(X_i\mid S_i,A_i,Z_i,,\;\boldsymbol{\zeta}) \qquad (3)$$

A key feature of our graphical model is that it assumes the phylogenetic model relating the outgroup data, O_i, and the deep ancestral allele, Z_i, is independent of the selection class, S_i. This is a relatively standard simplifying assumption that allows us to fit a single phylogenetic model per genomic block and use that model to define the distribution over the deep ancestral state Z_i in neutral sites as well as sites under selection. This implies that fitting of the outgroup phylogenetic scales, $\boldsymbol{\lambda}^{\mathbf{O}}$, as well as computation of the ancestral state priors, $\{P(Z_i \mid O_i,\ \hat{\lambda}^O_b)\}_{i \in E \cup F}$, can be performed prior to inference of selection (see **Parameter Inference**). The remaining terms in Equation (3) reflect our model for P&D. The intermediate ancestral allele, A_i, representing the most recent common ancestor (MRCA) of the target population in the local genealogy, allows us to separate the divergence portion of the model, $P(A_i \mid S_i, Z_i,\ \boldsymbol{\zeta})$, from the polymorphism portion, represented by $P(X_i|\ S_i, A_i, Z_i,\ \boldsymbol{\zeta})$.

The model for divergence is derived by assuming a divergence rate of $\lambda_b t$ for neutral sites and $\eta\lambda_b t$ for sites under selection, where t is the time since divergence, λ_b is the block-specific neutral divergence scale, and η is the relative divergence rate for sites under selection; η may be driven downward by negative selection or upward by positive selection (so it may be greater or less than 1). Any appropriate DNA substitution model can be used to obtain expressions for the probability of divergence. However, if the time since divergence is sufficiently short (as we typically expect), the following simple expressions provide a good approximation:

$$P(A_i = a \mid S_i = s, Z_i = z, \boldsymbol{\zeta}) \ = \ \begin{cases} \frac{1}{3}\lambda_b t & s = \text{neut},\ a \neq z \\ 1 - \lambda_b t & s = \text{neut},\ a = z \\ \frac{1}{3}\eta\lambda_b t & s = \text{sel},\ a \neq z \\ 1 - \eta\lambda_b t & s = \text{sel},\ a = z \end{cases} \tag{4}$$

The model for polymorphism is slightly more complicated. We assume an infinite sites model for the time since the MRCA (which is typically much shorter than the time since divergence, t), implying that $A_i \in \{X_i^{maj}, X_i^{min}\}$. The neutral polymorphism rate is given by the block-specific parameter $\theta_b = 4N_b\mu_b$, which captures both the local mutation rate and the total branch length of the local genealogy. Note that by considering the total branch length of the local genealogy, it also captures the effects of linked selection from nearby genes (i.e., background selection or hitchhiking). Direct selection on the elements of interest is modeled separately, by assuming a polymorphism rate of $\gamma\theta_b$ for sites under selection. Since polymorphisms should be less frequent in selected sites, we typically expect, but do not restrict $\gamma \leq 1$. Polymorphism rates are translated into probabilities of observing a polymorphic site using the factor a_n introduced by Watterson (1975) in his estimator for the effective population size: $a_n = \sum_{k=1}^{n-1} 1/k$, where n is the number of samples in the target population. Note that if the number of sampled genomes, n, is constant across all sites, the factor a_n serves as a constant scaling factor and its value is of no real consequence to the model. However, this approach can allow our model to accommodate small amounts of missing data by associating site i with the factor a_{n_i} that corresponds to the number of samples n_i with sequence data at that site (see **Discussion**).

The neutral processes for polymorphism and divergence are assumed to be independent, implying that X_i and Z_i are conditionally independent given A_i and $S_i = \text{neut}$. The derived allele in a polymorphic site is chosen uniformly at random among the three bases that are different from A_i, and its frequency is chosen to be in one of the three intervals $(0, f)$, $[f, 1-f]$, or $(1-f, 1)$ with probabilities β_1, β_2, and β_3, respectively. The distinction between the two high-frequency classes, $[f, 1-f]$ and $(1-f, 1)$, is required since they result in different *minor* allele frequency classes: $Y_i = \text{H}$ and $Y_i = \text{L}$, respectively. Note that sites with $Y_i = \text{H}$ are known to have a high derived allele frequency (since both alleles have high frequencies), but sites with $Y_i = \text{L}$ can have either low or "very high" derived allele frequency, depending on the identity of the derived allele.

$$P\left(X_i = (x^{maj}, x^{min}, y) \middle|\, S_i = \text{neut}, A_i = a,\ \boldsymbol{\zeta}\right) = \qquad (5)$$

$$\begin{cases} 1 - \theta_b a_n & y = \text{M},\ a = x^{maj} \\ \frac{1}{3}\beta_1\theta_b a_n & y = \text{L}\ ,\ a = x^{maj} \\ \frac{1}{3}\beta_3\theta_b a_n & y = \text{L}\ ,\ a = x^{min} \\ \frac{1}{3}\beta_2\theta_b a_n & y = \text{H}\ ,\ a \in \{x^{maj}, x^{min}\} \\ 0 & \text{otherwise} \end{cases}$$

For sites under selection, our model has to induce a somewhat artificial dependence between the polymorphism and divergence processes, stemming from the fact that polymorphisms are contributed only by sites under weak negative selection, and divergences are generated only in positively selected sites. Thus a selected site may not experience both polymorphism and divergence. This complication is alleviated by the fact that selected polymorphisms are restricted to low DAF, implying that $Y_i = \mathrm{L}$ and $X_i^{maj} = A_i$. As with neutral polymorphisms, the derived allele is chosen uniformly at random among the three bases that are different from A_i.

$$P\left(X_i = (x^{maj}, x^{min}, y) \middle|\, S_i = \text{sel}, A_i = a, Z_i = z,\ \boldsymbol{\zeta}\right) = \tag{6}$$
$$\begin{cases} 1 - \gamma\theta_b a_n & y = \mathrm{M},\ z = a = x^{maj} \\ 1 & y = \mathrm{M},\ z \neq a = x^{maj} \\ \frac{1}{3}\gamma\theta_b a_n & y = \mathrm{L},\ z = a = x^{maj} \\ 0 & \text{otherwise} \end{cases}$$

The P&D model can be summarized using a single conditional distribution table, $P(X_i \mid Z_i, S_i, \boldsymbol{\zeta})$ (see Table 3), obtained by combining Equations (4–6) and integrating over $A_i \in \{X_i^{maj}, X_i^{min}\}$.

Parameter Inference

The main objective of the inference procedure is estimation of the selection parameters, ρ, η, and γ. However, this requires that the neutral parameters, $\boldsymbol{\zeta}_{\text{neut}} = \left(\boldsymbol{\lambda}^{\mathbf{O}}, \boldsymbol{\lambda}, \boldsymbol{\theta}, \boldsymbol{\beta}\right)$, be inferred as well. In principle, it is possible to obtain maximum likelihood estimates (MLEs) for all parameters by joint inference using a single expectation-maximization (EM) algorithm. However, such an algorithm would be highly inefficient and thus impractical with the scale of genome-wide data, where the collection of functional elements can be expected to span millions of sites. Fortunately, several properties of the data and the probabilistic model enable an efficient method that produces a good approximation for the MLEs. First of all, notice that the flanking sites in F are expected to significantly outnumber the element sites in E (for short elements it is reasonable to expect $|F| > 100|E|$). This means that the neutral parameters, which contribute to the site-wise likelihood of both types of sites, can be inferred to a good approximation by maximizing the portion of the likelihood function that depends only on sequence data in flanking sites (see Equation (1)):

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\lambda}^{\mathbf{O}}, \boldsymbol{\lambda}, \boldsymbol{\theta}\,; \mathbf{X}_F, \mathbf{O}_F) \;=\; \prod_{b \in B} \prod_{i \in F_b} P(X_i \mid S_i = \text{neut},\ O_i,\ \lambda_b^O, \lambda_b, \theta_b \boldsymbol{\beta}) \,. \tag{7}$$

Conditioning on these fixed estimates of the neutral parameters, $\hat{\boldsymbol{\zeta}}_{\text{neut}}$, a straightforward EM algorithm can be employed to obtain MLEs for the selection parameters ρ, η, and γ. Since the selection parameters do not contribute to the site-wise likelihood at flanking sites, the portion of the likelihood function relevant to this maximization task depends only on sequence data in element sites:

$$\mathcal{L}(\rho, \eta, \gamma\,; \mathbf{X}_E, \mathbf{O}_E, \hat{\boldsymbol{\zeta}}_{\text{neut}}) \;=\; \prod_{b \in B} \prod_{i \in E_b} P(X_i \mid O_i,\ \rho, \eta, \gamma,\ \hat{\boldsymbol{\zeta}}_{\text{neut}}) \,. \tag{8}$$

An additional simplification in the inference procedure results from the observation that polymorphisms are typically rare (less than 1% of sites). This implies that the neutral divergence scales, $\boldsymbol{\lambda}$, can be estimated to a good approximation using only sequence data in monomorphic sites, decoupling this tasks from the task of inferring $\boldsymbol{\beta}$ and allowing separate estimation for each block-specific λ_b. Note that obtaining the exact MLEs of $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ requires joint estimation across all blocks. The complete inference procedure consists of the three separate stages summarized below (see Supplementary Methods for full description).

Phylogenetic Model Fitting. This stage produces estimates of the phylogenetic scaling factors $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}^{\mathbf{O}}$ and computes the prior distribution for the deep ancestral state, $P(Z_i \mid O_i, \boldsymbol{\lambda}^{\mathbf{O}})$, for each site $i \in F \cup E$. These tasks are carried out separately for each genomic block $b \in B$. The scaling factors are fitted to the alignment of the outgroup genomes using a user-supplied phylogenetic tree with branch lengths: λ_b is the scale applied to the branch leading to the target population, and λ_b^O is the scale applied to the remaining branches of the tree. As mentioned above, the estimation of λ_b makes use only of monomorphic sites in F_b, whereas λ_b^O is inferred using the entire set of putative neutral sites, F_b. This "neutral" outgroup scaling factor, $\hat{\lambda}_b^O$, is then used to compute the ancestral state prior $p(Z_i) \equiv P(Z_i \mid O_i, \hat{\lambda}_b^O)$ for all sites (neutral and not) in that block. This reflects the conditional independence assumed by our graphical model between joint random variable (O_i, Z_i) and the selection class S_i (Fig. 2). Note that the estimated outgroup scaling factors, $\hat{\boldsymbol{\lambda}}^{\mathbf{O}}$, can be discarded at this point, as they are not required by subsequent stages of the inference procedure. Both the scale-fitting step and the computation of ancestral priors are carried out using statistical phylogenetic procedures implemented in RPHAST (Hubisz, Pollard and Siepel, 2011).

Neutral Polymorphism Model Fitting. This stage produces estimates for the model parameters, $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, which describe the behaviour of neutral polymorphisms. MLEs of the polymorphism rate parameters, $\hat{\boldsymbol{\theta}}$, are relatively simple to obtain, as each $\hat{\theta}_b$ depends only on the observed polymorphism class Y_i across all sites in genomic block b. In the absence of missing data, the MLE has a simple closed-form solution, and in the presence of missing data, it can be computed using a simple numerical optimization procedure (see Supplementary Methods). The MLE for β_2 also has a simple closed-form solution. However, due to ancestral uncertainty, β_1 and β_3 need to be inferred using a simple EM algorithm that conditions on the neutral divergence rates $\hat{\boldsymbol{\lambda}}$ and ancestral state priors $\{p(Z_i)\}_{i \in F}$ estimated in the previous stage.

Selection Inference. This is the main inference stage, in which the selection parameters ρ, η, and γ are estimated via an EM algorithm, conditioning on the estimates of the neutral parameters, $\hat{\boldsymbol{\zeta}}_{\text{neut}}$, obtained in the previous stages. The objective of this EM algorithm is to maximize the likelihood function $\mathcal{L}(\rho, \eta, \gamma\ ; \mathbf{X}_E, \mathbf{O}_E, \hat{\boldsymbol{\zeta}}_{\text{neut}})$ expressed in Equation (8). Assuming completely observed model variables, the log-likelihood function can be expressed as a function of counts of various site types defined by configurations of the model variables (observed and/or hidden). For a set of sites, Q, and a configuration of model variables, $\mathcal{X}$, we denote by $c_Q(\mathcal{X})$ the number of sites in Q that have configuration $\mathcal{X}$. Using this notation, and assuming completely observed model variables, the log-likelihood function can be expressed as follows (where C represents a term that does not depend on ρ, η, or γ):

$$\begin{aligned}
&\mathcal{L}(\rho, \eta, \gamma\,; \mathbf{X}, \mathbf{O}, \hat{\zeta}_{\text{neut}}) = \qquad (9)\\
&\quad c_E(S_i = \text{sel}) \ln(\rho) \;+\; c_E(S_i = \text{neut}) \ln(1-\rho) \;+\\
&\quad c_E(S_i = \text{sel}, Z_i \neq X_i^{maj}) \ln(\eta) \;+\; \sum_{b \in B} c_{E_b}(S_i = \text{sel}, Z_i = X_i^{maj}) \ln(1 - \eta\lambda_b t) \;+\\
&\quad c_E(S_i = \text{sel}, Y_i = \text{L}) \ln(\gamma) \;+\; \sum_{b \in B} c_{E_b}(S_i = \text{sel}, Y_i = \text{M}, Z_i = X_i^{maj}) \ln(1 - \gamma\theta_b a_n) \;+\; C\,,
\end{aligned}$$

Because the model variables Z_i and S_i are not observed, counts associated with them are treated as random variables, and an EM algorithm is employed to obtain the MLEs. Each iteration of the EM algorithm assumes an assignment to all model parameters, uses that assignment to compute the expectations of the counts involved in Equation (9) (E step), and then uses these expected counts to update the selection parameters ρ, η, and γ to values that maximize the expected log-likelihood function (M step). The expected counts associated with the E step are obtained by computing for each site $i \in E$ the following three posterior probabilities: $p(S_i = \text{neut})$, $p(S_i = \text{sel}, Z_i = X_i^{maj})$, and $p(S_i = \text{sel}, Z_i \neq X_i^{maj})$. The computation of these posteriors makes use of the expressions in Table 3 as well as the ancestral priors, $\{p(Z_i)\}$, computed in the phylogenetic model fitting stage. In the M step of each iteration, ρ is updated using a simple formula, however, the updates of η and γ require a simple numerical optimization procedure to be applied (see Supplementary Methods). The E and M steps are applied iteratively, until the values of the likelihood function reach convergence.

Extracting Information about the Mode of Selection from Parameter Estimates

After fitting the model to data, we will often be interested in deriving various properties of the selective forces acting on the functional elements. The estimated value of the parameter ρ has a clear interpretation as a measure of the extent to which sites in the elements are affected by selection (positive, weak negative, or strong negative). The two other selection parameters, η and γ, are more difficult to interpret. However, given a joint assignment to all model variables, it is possible to produce posterior expectations for various measurements that directly relate to the the different modes of selection. A useful measure for the extent to which positive selection has affected the collection of functional elements is, D_{p}, the number of fixed differences (from the ancestral state, Z) within element sites that are driven by positive selection (also referred to as the number of adaptive substitutions). A similar measurement pertaining to weak negative selection is P_{w}, the number of polymorphic sites subject to weakly negative selection. Expected values for D_{p} and P_{w} are obtained by summing over site-wise posterior probabilities, as in the E step of the EM algorithm for selection inference.

$$\begin{aligned}\mathbb{E}[D_{\text{p}}] &= \langle c_E(Y_i = \text{M}, Z_i \neq A_i, S_i = \text{sel})\rangle \qquad (10)\\ &= \sum_{i \in E\ |Y_i = \text{M}} P(Z_i \neq X_i^{maj}, S_i = \text{sel}\ |X_i, O_i,\ \boldsymbol{\zeta})\ ,\\ \mathbb{E}[P_{\text{w}}] &= \langle c_E(Y_i = \text{L}, S_i = \text{sel})\rangle \qquad (11)\\ &= \sum_{i \in E\ |Y_i = \text{L}} P(S_i = \text{sel}\ |X_i, O_i,\ \boldsymbol{\zeta})\ ,\end{aligned}$$

where $\langle c_E(\chi)\rangle$ denotes the expected number of element sites with model variable configuration χ (see previous section).

These formulas makes use of our two main assumptions regarding modes of selection, stating that divergence at selected sites occurs only due to positive selection, and polymorphism at selected sites occurs only due to weak negative selection and is restricted to 'L' sites. When comparing different collections of functional elements, we will typically be interested in versions of D_{p} and P_{w} normalized by the total number of element sites, $|E|$. Alternatively, by normalizing $\mathbb{E}[D_{\text{p}}]$ by the total (expected) number of divergences, we can also obtain an estimate of the fraction of fixed differences driven by positive selection, referred to in the literature as α (Smith and Eyre-Walker, 2002). Both measures, $\mathbb{E}[D_{\text{p}}]$ per site and α, provide useful information on the extent to which positive selection has influenced the functional elements of interest. While α has been used in several recent studies as a measure for positive selection (Smith and Eyre-Walker, 2002; Andolfatto, 2005), normalizing $\mathbb{E}[D_\text{p}]$ by the total number of sites has the advantage of being less sensitive to negative selection acting on the elements of interest (negatively selected sites will tend to reduce the overall number of divergences and thus increase α). In our analysis, whenever we refer to $\mathbb{E}[D_\text{p}]$ we either normalize it per site or, when specifically indicated, per 1,000 bp (kbp).

Confidence Intervals and Likelihood Ratio Tests

The full probabilistic nature of our model allows us not only to provide point estimates for values of interest, such as ρ, D_p, and P_w, but also to assess our confidence in these estimates and perform hypothesis testing. Approximate variances in parameter estimates can be derived from an estimated Fisher information matrix, using what is sometimes called the "curvature" method (Lehmann and Casella, 1998). Specifically, the Fisher information matrix is estimated by negating the 3×3 matrix of second derivatives (Hessian) of the log-likelihood function w.r.t. ρ, η, and γ evaluated at the estimated MLE. The Hessian can be estimated analytically for our model (see Supplementary Methods). The Fisher information matrix is then inverted to estimate the variance/covariance matrix for the three selection parameters, and confidence intervals for these parameters are obtained by taking the square root of the diagonal elements of this matrix.

To propagate variances in parameter estimates through to calculations of the posterior expected number of divergences under positive substitutions, we use the approximation $\mathbb{E}[D_\text{p}] \approx \rho\eta\bar{\lambda}t$, where $\bar{\lambda}$ is a weighted average of all λ_b values. Using a first-order Taylor approximation, we then estimate the variance of $\mathbb{E}[A]$ to be:

$$\text{Var}[\ \mathbb{E}[D_\text{p}]\ ] \approx (\eta\bar{\lambda}t)^2\text{Var}[\rho] + (\rho\bar{\lambda}t)^2\text{Var}[\eta] + 2(\eta\bar{\lambda}t)(\rho\bar{\lambda}t)\text{Cov}[\rho,\eta]. \quad (12)$$

The variance and covariance terms in this expression are extracted from the inverted Fisher information matrix. The variance of $\mathbb{E}[P_\text{w}]$ can be approximated by a parallel, but slightly more complex, calculation based on the approximation $\mathbb{E}[P_{\mathrm{w}}] \approx \rho\gamma a_n(\bar{\theta} - \eta\overline{\lambda\theta}t)$, where $\bar{\theta}$ is a weighted average of all θ_b values and $\overline{\lambda\theta}$ is a weighted average of all products $\lambda_b\theta_b$ (see Supplementary Methods). Note that these curvature-based estimates of the variance do not capture uncertainty in the estimates of the neutral parameters. However, uncertainty in the neutral estimates should be fairly low assuming a sufficient number of putative neutral sites within the relevant genomic blocks. This can be ensured by filtering element sites in genomic blocks with too few putative neutral sites.

The full likelihood framework enables testing the hypothesis that any of the selection parameters, ρ, η, or γ, is greater than zero. The null hypothesis of $\rho = 0$ is associated with no selection of any kind, whereas the null hypotheses of $\eta = 0$ and $\gamma = 0$ are associated with no positive and weak negative selection, respectively. These tests are performed using likelihood ratio tests (LRT), by fitting the model twice, once without restricting any of the parameters and once while fixing the parameter of interest to zero. Twice the difference between the log likelihoods associated with these estimates is then treated as a test statistic, in the usual way. In the case of η and γ this is a fairly straightforward LRT with one degree of freedom, except that the null hypothesis is at the boundary of the alternative hypothesis (because these parameters are bounded by zero). Therefore, the asymptotic distribution is equal to a 50:50 mixture of a χ^2 distribution with one degree of freedom and a point mass at zero in computing approximate p-values (Chernoff, 1954; Self and Liang, 1987). The case of ρ is more complex, because, in addition to the boundary issue above, a value of $\rho = 0$ causes η and γ to become irrelevant. In the LRTs we applied in the analysis presented in this paper, we took a conservative approach and used use a χ^2 distribution with three degrees of freedom for p-value calculations of the LRT for $\rho > 0$, and a χ^2 distribution with one degree of freedom for p-value calculations of the LRT for $\eta > 0$ and $\gamma > 0$. Control analysis we did using elements randomly selected from putative neutral regions of the genome indicate that this approach is slightly conservative but still has reasonable power (see **Results**).

Implementation and Software

INSIGHT software consists of several software components, each implementing a different stage of the inference procedure. The key component is the program that performs the EM algorithm for selection inference (as well as the EM algorithm for inferring β_1 and β_3). This program is implemented in C and provides maximum likelihood estimates of the model parameters ρ, η, and γ, and the posterior expected counts $\mathbb{E}[D_\text{p}]$ and $\mathbb{E}[P_\text{w}]$. It computes approximate standard errors for the estimated values using the curvature method and can optionally produce posterior probabilities over hidden model variables (S_i, Z_i, and A_i) across the input set of sites, which can be used to derive posterior expectations of other measures of interest (such as α). The program takes about one minute to analyze a collection of genomic elements spanning 300,000 bases . The phylogenetic model fitting stage is implemented separately using procedures from RPHAST (Hubisz, Pollard and Siepel, 2011), and additional scripts are used for processing and filtering the variation data. Source code, documentation and sample files are available for download from http://compgen.bscb.cornell.edu/INSIGHT/.

Simulations

In order to test our model and inference procedure, we generated a large collection of simulated elements under different scenarios of selection together with flanking neutral regions. Each simulation followed a population phylogeny that spanned three outgroup populations and four different target populations. The simulations were designed to reflect the joint evolutionary history of humans and their closest primate relatives: chimpanzee, orangutan, and rhesus macaque. The effective population size was held constant at N_e =10,000 across the outgroup portion of the phylogeny, and divergence times of 6.5, 17.5, and 25 million years ago were set for the chimpanzee, orangutan, and rhesus macaque outgroup populations, respectively. Each of the four target populations was simulated using a different demographic trajectory that allowed us to validate the robustness of our methods to different demographic scenarios (Supplementary Table S1). One target population was simulated with constant size since divergence from chimpanzee, one with a moderate population expansion, and the two others were simulated with population bottlenecks and exponential expansions (see Supplementary Methods). The intensity and timing of the bottlenecks and expansions were taken from the demographic model suggested by Gutenkunst et al. (2009), reflecting demographic histories of African, Eurpean, and East-Asian populations. In each simulation we sampled a single haploid genome from each of the three outgroup species and 50 diploid individuals from one of the target populations. This closely resembles the settings used in our human data analysis, which is based on a collection of 54 unrelated individual human genomes (see details below).

Each simulation included a 10 bp element and 5,000 neutral sites flanking it on each side (leading to a total of 10,010 sites per simulated block). The blocks were simulated separately, implying complete linkage equilibrium between the different elements. However, recombination was allowed within each simulated block at a constant population-scaled rate of $\rho = 4N_er = 4.4{\times}10^{-4}$ recombinations per nucleotide position. The population-scaled mutation rate, $\theta = 4N_e\mu$, was assumed to have a mean value of 7.2×10^{-4}, consistent with human population genomic studies, but was allowed to vary across the different simulated blocks. The specific rate at each block was sampled from a normal distribution having this mean and a standard deviation equal to one tenth of the mean. Each nucleotide position in each simulated block was assigned to one of four selection classes: neutral evolution ($2N_es = 0$), strong negative selection ($2N_es = -100$), weak negative selection ($2N_es = -10$), and positive selection ($2N_es = 10$). The 10 kb flanking sites were all assigned to the neutral class, and the 10 bp of each simulated element were assigned to one of the four classes using a multinomial distribution to determine the number of sites in each selection class. We used a range of multinomial distributions across the different collections of synthetic elements we analyzed.

The simulations were carried out using SFS_CODE (Hernandez, 2008), which provides a flexible framework for full forward simulation of sequence evolution in populations with selection. To express times in units of $2N_e$ generations, as required by SFS_CODE, we used the ancestral effective population size of 10,000 and assumed a generation time of 20 years. To save in computational cost, we used $N_{\text{sim}} = 1{,}000$ individuals in forward simulations. Notice that as long as N_{sim} is sufficiently large to limit sampling error, this strategy should have little effect on results, because all parameters are expressed in population-scaled form. We used the default "burn-in" of $5 \times 2N_{\text{sim}}$=10,000 generations before initiating the specified demographic scenario. Weak and strong negative selection were applied constantly across the phylogeny at desginated sites. Positive selection was handled in a slightly different way in order to reflect a scenario in which an element is initially under constraint but then comes under positive selection in the lineage leading to the target population and then stabilizes again once an advantageous allele reaches fixation. The default behavior in SFS_CODE models positive selection by assuming positively selected sites constantly carry a suboptimal allele, which tends to produce repeated fixation events. To ensure a more stable behavior in positively selected sites, these sites were defined to be under weak negative selection ($2N_es = -10$) in the ancestral and outgroup populations, but then switched to positive selection ($2N_es = 10$) on the human lineage immediately after the human/chimpanzee divergence. The site is then simulated under positive selection for more than 6 million years, and at point 300,000 years ago the site reverts to weak negative selection. This strategy ensures that positively selected sites have the opportunity of undergoing a selective sweep (although they are not guaranteed to do so), but prevents recurrent and ongoing positive selection in the present day population from obscuring the main signal of long term adaptation.

We tested INSIGHT on the simulated data by constructing from the initial pool of elements several collections of elements of size 10,000-20,000 with different proportions of sites under selection and different combinations of the different selective modes. The inference procedure of INSIGHT was applied separately to each of these collections, using the three stages of inference, as described above. The expected number of divergences driven by positive selection, $\mathbb{E}[D_{\text{p}}]$, and the expected number of polymorphisms under weak negative selection, $\mathbb{E}[P_{\text{w}}]$, were extracted using the MLEs of ρ, η, and γ. Standard errors were computed for estimates of the model parameters and expected counts using the curvature method. Finally, the estimated values of ρ, $\mathbb{E}[D_{\text{p}}]$, and $\mathbb{E}[P_{\text{w}}]$ were compared to the true values of these measures in simulation. The true value of ρ was simply the proportion of sites under selection across elements in the collection. The true value of D_{p} was the number of fixed differences (with respect to the true ancestral allele Z) that occured in positively selected sites. Since the distinction between weak and strong negative selection is somewhat arbitrary, for the true value of P_w we used the number of negatively selected sites (weak or strong) that are polymorphic. Positively selected polymorphic sites are not considered for this purpose (see **Results**).

Simple Site-count-based Estimates

As a comparison point for the model-based estimates from INSIGHT, we made use of estimators for the fraction of sites under selection (ρ) and the number of adaptive substitutions (D_p) based on simple counts of nucleotide substitutions and/or polymorphic sites. As a divergence-based estimator for ρ, we use a quantity introduced by Kondrashov and Crow (1993):

$$\hat{\rho}_\text{Div} \;=\; 1 - \frac{f_E}{f_F} \;=\; 1 - \frac{D_E\,|F|}{|E|\;D_F}, \tag{13}$$

where f_E and f_F are the rate at which divergences occur within the elements and within their neutral flanking regions, respectively, and the rate for a set of sites $X \in \{E, F\}$ is estimated by the total number of divergences in X, D_X, divided by the total number of sites in X. Similar to our model, this estimator makes simplifying assumptions about the process of site substitution that are justified for short evolutionary time scales. However, unlike our model, $\hat{\rho}_\text{Div}$ ignores the effect of positive selection on divergence and it pools counts across elements in a manner that does not account for variable mutation rates across the genome. As a simple polymorphism-based estimator we used an analogous quantity:

$$\hat{\rho}_\text{Poly} = \;=\; 1 - \frac{P_E\,|F|}{|E|\;P_F} \tag{14}$$

where P_X is the number of polymorphic sites in X. The main differences with respect to INSIGHT are that $\hat{\rho}_\text{Poly}$ assumes selected sites cannot be polymorphic (hence it poorly handles weak selection) and, like $\hat{\rho}_\text{Div}$, it naively pools counts across loci.

As a simple estimator for $\mathbb{E}[D_\text{p}]$, we used the estimator for α, the fraction of fixed differences under selection, introduced by Smith and Eyre-Walker (2002) based on the McDonald-Kreitman

(MK) test (McDonald and Kreitman, 1991):

$$\hat{D}_{\text{P-MK}} = D_E \, \hat{\alpha}_{\text{MK}} = D_E \left(1 - \frac{P_E \, D_F}{D_E \, P_F} \right) = D_E - \frac{P_E \, D_F}{P_F} \, . \tag{15}$$

Like $\hat{\rho}_{\text{Poly}}$, this estimator implicitly assumes no polymorphisms occur in selected sites, and like both simple estimators for ρ, it naively pool counts across loci.

Analysis of Human Transcription Factor Binding Sites

In order to test INSIGHT on real data, we applied it to transcription factor binding sites (TFBSs) identified in multiple human cell types using genome-wide chromatin immunoprecipitation and sequencing (ChIP-seq) experiments carried out by the ENCODE project (Myers et al., 2011). A comprehensive study of 78 transcription factors (TFs) using INSIGHT is reported in Arbiza et al. (2012), together with complete details on the data preparation pipeline. Here we briefly outline this pipeline and report in greater detail the analysis done for four of these TFs (see **Results**). Our pipeline for identifying high-confidence binding sites based on the available ChIP-seq data involved de novo motif discovery (using MEME; Bailey and Elkan, 1994) , manual inspection, and binding-site prediction at ChIP-seq peaks (using MAST; $p < 0.0001$, $E < 10$). We used ChIP-seq data from multiple cell lines, and merged binding sites across cell lines to generate a single collection of binding sites for each TF.

Information about human polymorphisms came from the "69 Genomes" data set from Complete Genomics (http://www.completegenomics.com/public-data/69-Genomes/). While larger data sets are available (The 1000 Genomes Project Consortium, 2010), this one was selected for its high coverage, which reduces the effect of genotyping error and allows singleton variants to be characterized with fairly high confidence. We identified 54 unrelated individuals in this set, by eliminating 13 individuals out of the 17-member CEPH pedigree and the children in the two trios. Genotype calls for these individuals were extracted from the 'masterVar' files. Outgroup data was obtained from the alignments in the UCSC Genome Browser of the chimpanzee (panTro2), orangutan (ponAbe2), and rhesus Macaque (rheMac2) genomes with the human reference genome (hg19). Filters were applied to eliminate repetitive sequences, recent duplications, CpG sites, CpG islands, and regions not showing conserved synteny with outgroup genomes. Our analysis considered the autosomes only (chromosomes 1–22). A collection of putative neutral sites was computed by excluding exons of known protein-coding and RNA genes and 1,000 bp flanking them on each side, as well as conserved noncoding elements and 100 bp flanking regions, leaving an average of 3,881 sites per 10 kbp block. Genomic blocks with less than 100 putative neutral sites were filtered, and binding sites in these blocks were removed from the analysis.

As a pre-processing step for INSIGHT inference on genomic data, we applied the phylogenetic model fitting stage genome-wide, using 10 kb sliding windows with 5kb overlap. The phylogenetic scales λ_b and λ_b^O were fitted to the outgroup alignment in each 10 kb window and then associated with the 5 kb block at the center of the bigger block. The outgroup scaling factor, λ_b^O, was used to compute priors for the ancestral state Z_i for all sites in the 5 kb block, and the divergence scale, λ, was later used as the neutral divergence rate for all sites in the 5 kb block. The phylogenetic model was fitted by assuming the standard topology, relative branch-lengths, and substitution rate matrix assumed for the four-species primate phylogeny. Estimates of neutral polymorphism rates, $\boldsymbol{\theta}$, were also obtained genome-wide using the same sliding window approach. We then use these pre-computed estimates when performing selection inference on different collections of elements. For each nucleotide spanned by a given set of elements, we extract the ancestral prior, as well as the estimates of λ_b and θ_b associated with the 5 kb block to which this nucleotide belongs. We then infer the neutral polymorphism class proportions $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)$ using neutral sites in the blocks flanking the elements in the collection, and conclude by executing the EM algorithm for selection inference. This approach of generating a "neutral field" by processing information from putative neutral sites allowed us to separate the time consuming stages of the inference procedure from the main stage of selection inference, leading to a much more streamlined analysis.

Results

Simulations

In order to study the accuracy of our inference method, we applied it to various collections of elements simulated with their flanks using SFS_CODE (see **Methods**). We started by considering four collections, each with 20,000 elements, representing distinct scenarios of selection characterized by different proportions of sites in each of the four categories: neutral, weak negative (WN), strong negative (SN), and strong positive (SP) (Fig. 3A). In order to test the sensitivity of INSIGHT to weak signal, we constructed two collections (gray and red) with low proportions of sites under selection (0.05 and 0.27, respectively). In both cases, the true value of ρ was within one standard error of the estimated values (standard errors approximated using the curvature method; see **Methods**). The small proportions of WN sites in these collections resulted in small numbers of polymorphisms under (weak) negative selection, P_{w}, which were slightly over-estimated by the expected posterior values, but within one standard error. The main factor distinguishing the second collection from the first was a larger number of SP sites, resulting in a higher number of adaptive divergences, D_{p}. For both collections, the posterior expected values, $\mathbb{E}[D_{\mathrm{p}}]$, were only slightly lower than the true values (in about 0.3 divergences per kbp in both cases). The other two collections (blue and green; Fig. 3A) had higher proportions of sites under selection (0.66 and 0.85, respectively), which allowed for somewhat greater power. This increase in power was apparent both by a smaller deviation of estimates from the true values and a reduction of about half in the standard errors when comparing the second pair of data sets with the first pair. The number of deleterious polymorphisms, P_{w}, was estimated accurately in the third collection (blue) and over estimated very slightly in the fourth collection (green) simulated with positive selection. Estimates of the number of adaptive divergences, D_{p}, were highly accurate in both data sets.

We compared our model-based estimates with the simple site-count-based estimates for ρ and D_{p} (see **Methods**). The simple divergence-based estimator, $\hat{\rho}_{\mathrm{Div}}$, accurately estimated the fraction of sites under selection in the two collections with low or no positive selection (gray and blue), and grossly under-estimated it in the other two collections ($\hat{\rho}_{\text{Div}}$ was -0.52 in the second collection). This is, perhaps, not surprising, since this estimate ignores the effect of positive selection on the divergence rate. In particular, when the divergence rate is higher in the elements compared to the neutral flanks, this estimator produces negative values. The polymorphism-based estimator, $\hat{\rho}_{\text{Poly}}$, showed better correlation with the true values of ρ, however, it consistently under-estimated the true fraction of sites under selection by 0.05-0.22 due to the presence of deleterious polymorphisms. Even the small proportion of WN sites in the first two collections lead to a non-negligible under-estimation of ρ (by 0.05 and 0.09 resp.). The simple MK-based estimator for the number of adaptive divergences, $\hat{D}_{\text{p-MK}}$, also showed a similar sensitivity to deleterious polymorphisms, which resulted in under-estimation of 0.6-1.1 divergences per kbp. We note that it is possible to obtain slightly better polymorphism-based estimates by ignoring low-frequency polymorphisms, but even this correction does not eliminate the under-estimation completely (see discussion in next section).

In order to test the potential effects of the demographic history of the target population on our inference, we performed inference on samples taken from four different target populations simulated under different demographic histories: one with constant population size since divergence from chimpanzee, one with a moderate population expansion, and two other with a severe population bottleneck followed by exponential expansion (Supplementary Table S1). We analyzed seven collections of 10,000 elements simulated in each of these populations, with true values of ρ ranging from 0.4-1.0, keeping the proportion within selected sites constant at 45% WN, 50% SN, and 5% PD. Inference was performed separately for each of the 7×4 data set, and the inferred values were compared to the true values in simulation (Fig. 3B).

The model-based estimates of ρ did not show any bias across the different demographic scenarios, although the estimation error was slightly higher for data simulated with population bottlenecks. The divergence-based estimates, $\hat{\rho}_{\text{Div}}$, were poor across all simulations due to the effects of positive selection, as mentioned above. The polymorphism-based estimates, $\hat{\rho}_{\text{Poly}}$, consistently under-estimated the true values, with an average relative decrease of 24% in the first two demographic scenarios and an average relative decrease of 42% in the scenarios that contained bottlenecks. Similar under-estimation of D_{p} was also observed for the MK-based estimator $\hat{D}_{\text{p-MK}}$. In contrast, the posterior expected number of positively selected divergences, $\mathbb{E}[D_{\mathrm{p}}]$, provided an unbiased estimate of D_{p}, with higher estimation error for data simulated with population bottlenecks, as observed for ρ. Estimates of the number of polymorphisms under selection, P_{w}, appeared to be more sensitive to demography. This is, perhaps, not a surprising outcome considering the fact that demography directly affects the site frequency spectrum. Nevertheless, it is reassuring to see that this sensitivity does not propagate to estimates of the other parameters.

A key feature of the model is the distinction between high- and low-frequency polymorphisms. In the above analysis, we made this distinction using the default threshold of 15%, which was shown by previous studies to be a reasonable upper bound on the DAF of weakly negative polymorphic sites (Charlesworth and Eyre-Walker, 2008). However, since this threshold is the only component of the model that is pre-determined by the user and not fitted to data, it is important to understand the way it affects the inferred parameter estimates. In order to test the effect of the frequency threshold on parameter estimates, we performed inference on the same seven collections mentioned above using different frequency thresholds ranging from 1% to 40% (Fig. 3C). Setting a frequency threshold that is too low (f <7%) resulted in under-estimation of ρ, D_{p} and P_{w}, due to the presence of selected polymorphisms with DAF$>f$. Frequency thresholds above 7% did not lead to under-estimation, and the estimation error for ρ did not fluctuate much when f was in the range 7-20%. The standard error of our estimates showed a moderate, but steady, increase with f, owing to the direct effect of the number of high frequency polymorphisms on power. We conclude that a frequency threshold of 15% provides a reasonably conservative cutoff that does not compromise the statistical power of our method.

An important feature of the model for selection implemented in SFS_CODE is that it is based on full forward simulation of a population, and as such, does not necessarily adhere to the basic assumptions made by our probabilistic model. This allowed us to use the simulated data to validate and revisit these assumptions. One assumption we make, which is critical in our interpretation of

$\mathbb{E}[D_\mathrm{p}]$ as a measure of positive selection, is that sites under negative selection (weak or strong) do not contribute to divergence. Indeed, throughout all our simulations, we find that mutations in negatively selected sites never reach fixation. Polymorphisms in selected sites, on the other hand, do not strictly follow our formal assumptions restricting them to WN sites. Across all simulated elements, 70% of polymorphisms under selection occurred in WN sites, 22% occurred in SN sites and 8% occurred in positively selected sites. Since the distinction between weak and strong negative selection is somewhat arbitrary, low-frequency polymorphisms in sites under strong negative selection do not constitute a major violation of our assumptions. Low-frequency polymorphisms under positive selection are potentially more problematic because they should drive the estimates of P_w upward. In practice, however, our experiments using relatively high proportions of positive selection (red and green in Fig. 3A) show only a slight excess in $\mathbb{E}[P_\mathrm{w}]$ compared to the true value of P_w. The modeling assumption whose violation would be most problematic for inference is the presence of high-frequency polymorphisms under selection. This is because our model explicitly assumes high-frequency polymorphisms are neutral, and thus an excess in such sites will drive the estimate of ρ downward. In our simulated data, we found that when using the default frequency threshold of 15% to distinguish between high hand low frequencies, 98.8% of high-frequency polymorphisms in elements were generated in neutral sites, 0.7% in WN sites and 0.5% in SP sites. The high level of accuracy in the estimation of ρ by INSIGHT across all simulated data indicates that our model is robust to this small fraction of non-neutral high frequency polymorphisms.

Analysis of Human Transcription Factor Binding Sites

We applied INSIGHT to several collections of transcription factor binding sites (TFBSs) identified in human cell lines using ChIP-seq assays. A comprehensive genome-wide study of selection in human TFBSs based on this analysis is reported in Arbiza et al. (2012). Here we focus in greater depth on a small subset of transcription factors and demonstrate usage of INSIGHT in a genome-wide setting. Before turning to analyze the TFBS data, we performed a brief control study using collections of short regions extracted from neutral regions genome-wide. For this purpose we used the same collection of genomic positions used for identifying putative neutral sites in the flanking regions in our analysis (see **Methods**). We constructed 500 mutually exclusive collections of "neutral" elements, each collection with 28,350 elements, 10 bp long, spaced roughly 50,000 bp apart. The values of ρ estimated by INSIGHT for these 500 collections were fairly low, with a median of 0.04, and maximum of 0.23 (Fig. 4A; left). We also recorded the test statistic for the hypothesis that $\rho > 0$ for each collection and compared the distribution of statistics to a χ^2 distribution with three degrees of freedom (Fig. 4A; right). Overall, the shape of the χ^2 distribution appeared to fit the distribution of test statistics quite well. Testing specific cutoff points, we found that that eight collections (1.6%) showed statistics that exceeded the $p = 0.01$ cutoff, and 23 collections (4.6%) showed statistics that exceeded the $p = 0.05$ cutoff, indicating that this is an adequate approach for significance testing.

We turned next to separately analyze the collections of binding sites of four transcription factors (TFs): BRCA1, CTCF, GATA2, and SUZ12 (Table 4). We found significant evidence of selection, and weak negative selection in particular, in the binding sites of all four TFs ($p = 0.01$), with estimated values of ρ ranging from 0.32–0.43. GATA2 and SUZ12 both showed elevated estimates of D_{p}, at 1.11 and 0.86 adaptive divergences per kbp, respectively, but only the estimate for GATA2 was found to be significantly greater than 0 ($p = 0.001$). As in our simulation study, we used the default setting of 15% for the frequency threshold distinguishing low- and high-frequency polymorphisms (see above). In order to verify robustness of our estimates to this threshold, we repeated our inference procedure for the four TFs using a range thresholds (Fig. 4B). The estimates showed little sensitivity to fluctuations of the frequency threshold within the interval 10-20%, with TFs that have sparser data (such as SUZ12) showing slightly higher sensitivity.

For comparison, we computed the simple site-count-based estimates, $\hat{\rho}_{\text{Poly}}$ and $\hat{\rho}_{\text{Div}}$, for each of the four TFs (Fig. 4B). The divergence-based estimates were slightly higher than our model-based estimates for CTCF and BRCA1 (0.38 and 0.48, resp.), and much lower than our model-based estimates for SUZ12 and GATA2 (0.18 and 0.06, respectively), due to signatures of positive selection found in the binding sites of these TFs. Similarly, the estimates of ρ produced by INSIGHT were more than twice higher than $\hat{\rho}_{\text{Poly}}$ for all four TFs, due to the presence of weak negative selection in these binding sites. It is worth noting that the simple polymorphism-based estimates of ρ can be made to better accommodate weak negative selection by redefining P_E and P_F in Equation (14) as the number of high-frequency polymorphisms in the binding sites and flanks, respectively. In the case of SUZ12, these corrected polymorphism-based estimates are very close to our model-based estimates. On the other hand, for other TFs, such as CTCF, the correction appears to be only partially effective, as it does not consider divergence patterns and has the basic pitfall that it does not adequately account for variation in mutation rate and genealogical background along the genome.

When studying selection on functional elements that have a certain well-defined functional structure, it is possible to use INSIGHT to characterize the different selective forces that act on different components of the element, by performing separate inference on each of these components. In the case of binding sites, each TF is associated with a sequence motif that defines its binding preference, and each nucleotide in a binding site of that TF is uniquely mapped to a position in that motif. We can thus partition all nucleotides in binding sites of a given TF based on the motif position to which they are mapped and then infer selection using INSIGHT separately for each position along the motif. We carried out a position-specific analysis for the binding sites of GATA2, which is a TF shown in our initial analysis to have a significant signal for positive selection. For each of the 11 positions defined by that motif (prior to trimming), we inferred ρ, $\mathbb{E}[D_{\text{p}}]$, and $\mathbb{E}[P_{\text{w}}]$ (Fig. 4C). Five out of the seven positions analyzed in our initial analysis of GATA2 showed significant signatures of selection ($\rho > 0$; $p = 0.01$). Positions 6 and 7, show a significant signature of positive selection ($\eta > 0$; $p = 0.01$), with an estimated expected number of adaptive divergences of 2.9 and 1.8 divergences per kbp, respectively. Furthermore, positions 6 and 8 show significant evidence of weak negative selection ($p = 0.01$).

[ILAN: THIS LAST PARA IS TOO WORDY. NEEDS A STRONGER BOTTOM LINE]

Another commonly used measure of position-specific selective pressure for TFBSs is the information content (IC) associated with the motif at each position. The IC of a given motif position indicates the level of sequence agreement among nucleotides mapped to that position across all instances of the binding site found along the genome. One might therefore expect IC to be correlated with the selective pressure acting along the motif. When comparing IC with our position-specific estimates of ρ for GATA2, we find a coarse-grained correlation indicated by the fact that five of the seven positions with IC $> \frac{1}{2}$ and none of the four position with IC $\leq \frac{1}{2}$ have values of ρ that are significantly greater than 0 (Fig. 4C). Interestingly, we do not find high correlation between IC and the estimated values of ρ at positions with IC $> \frac{1}{2}$. This is perhaps not surprising, since ρ and IC measure two different aspects of sequence conservation in TFBSs. IC measures conservation by directly comparing many different genome-wide instances of the binding site to each other and using only the human reference genome sequence. The INSIGHT estimates of ρ are obtained by aggregating the separate patterns of sequence variation and divergence of each genomic position without directly comparing the sequence of different instances of the binding site to each other.

Discussion

In recent years, methods based on evolutionary conservation have become a primary tool for identifying and characterizing noncoding functional elements (Margulies et al., 2003; Siepel et al., 2005; Cooper et al., 2005; Pollard et al., 2010), but these methods are limited by their consideration of relatively long evolutionary time scales and their reliance on multiple alignments of sometimes quite divergent sequences. They could potentially be made more sensitive to lineage-specific evolutionary patterns by incorporating newly available data sets describing population genomic variation, together with comparative genomic data for closely related species. However, a strategy of this kind requires a means for efficiently pooling information from many genomic sites, because data describing variation within populations and divergence on short time scales is necessarily sparse. This is the motivation for the Inference of Natural Selection from Interspersed Genomically coHerent elemenTs (INSIGHT) method introduced here. INSIGHT bears some similarities to McDonald-Kreitman (MK)-based methods (McDonald and Kreitman, 1991; Smith and Eyre-Walker, 2002;

Bierne and Eyre-Walker, 2004; Andolfatto, 2005), Poisson Random Field (PRF) methods (Sawyer and Hartl, 1992; Bustamante et al., 2002, 2005; Williamson et al., 2005), and related methods for characterizing the distribution of fitness effects (Eyre-Walker, Woolfit and Phelps, 2006; Boyko et al., 2008; Eyre-Walker and Keightley, 2009), but it differs from previous methods in several important respects. In particular, unlike MK-methods, it is based on a full generative probabilistic model, it pools information from many loci in a manner that accommodates variation in neutral rates and genealogical backgrounds, and it explicitly models weak negative selection. Unlike PRF methods, it directly describes contrasting patterns of polymorphism and divergence in elements of interest with flanking sites to mitigate biases from complex demographic histories. Because it is based on a fully generative probabilistic model, INSIGHT allows for straightforward likelihood ratio tests of various hypotheses of interest, and it allows parameter variances to be approximately characterized using standard methods. For these reasons, we expect INSIGHT to be a valuable complement both to existing methods for analyzing noncoding regions based on long-term evolutionary conservation, and to methods for analyzing protein-coding sequences based on patterns of polymorphism and divergence.

Our relatively simple probabilistic model is specifically designed to exploit newly available genome-scale data sets describing both candidate functional elements (Myers et al., 2011; Gerstein et al., 2010; Roy et al., 2010) and variation within populations (The 1000 Genomes Project Consortium, 2010; Mackay et al., 2012). Nevertheless, a naive approach to parameter estimation would still be prohibitively CPU-intensive for a typical genome-wide data set. We achieve major gains in efficiency by decomposing the inference procedure into three separate steps, concerned with the estimation of the phylogenetic, neutral, and selection parameters, respectively. This decomposition relies on the simplifying assumption that neutral sites within the elements of interest contain negligible information about the neutral parameters of the model, because they are vastly outnumbered by the flanking neutral sites—a property that can typically be guaranteeed by construction. It also depends on the use of a single phylogenetic model per locus in estimating the prior distribution of the ancestral allele at all sites, which should be adequate as long as the branch lengths of the phylogeny are not too long. Notably, the first two of these steps can be performed in preprocessing and reused in the analysis of any set of loci that use the same flanking regions. If desired, the neutral flanks can be designed to maximize the potential for reuse, for example, by tiling the genome with 10 kbp blocks and associating each element with the neutral sites of the nearest block. This strategy would allow the neutral and phylogenetic parameters to be pre-estimated in each block and reused in any number of subsequent analyses. Importantly, these steps dominate the running time (particularly the phylogenetic estimation step). The final stage, in which the parameters ρ, η, and γ are estimated, is independent of the number of genomes considered and typically takes less than a minute.

It is worth emphasizing that INSIGHT can be applied to any collection of genomic elements, provided each one is sufficiently short that it does not span regions having markedly different mutation rates or genealogies, and provided each element can be associated with nearby sites likely to be free from the effects of selection. In this paper, we have focused on the specific instance of a genome-wide set of transcription factor binding sites (TFBSs) for a particular transcription factor, as identified by ChIP-seq, but many other types of analysis are possible. For example, in related work (Arbiza et al., 2012), we have examined various subsets of TFBSs, such as those associated with genes of a particular Gene Ontology category or expressed at a various levels, and those having various levels of predicted binding affinity. Similar analyses could be performed with other short genomic intervals, such as the DNA templates for the binding sites of microRNAs or RNA-binding proteins. As demonstrated in the position-specific analysis reported here, the method can also be applied to well-defined subsets of positions within elements. In this case, our analysis provided a view of the position-specific influence of natural selection on binding sites that complements the information captured by the position weight matrix of the motif, indicating, among other things, a significant influence from positive selection in two positions of the GATA2 TFBSs. Similar analyses could be used to contrast sequences corresponding to different regions of protein or RNA structures, regions of the genome having different epigenomic marks, sex chromosomes and autosomes, or any number of other biologically significant genomic partitions.

INSIGHT could be extended in various ways to improve model fit and broaden its utility. In this analysis we had a sufficiently large and complete collection of human variation data to simply discard positions with missing data in one or more samples. In cases of more missing data, however, it may be worthwhile to use the strategy of adjusting Watterson's constant a_n in the appropriate conditional distributions (see Table 3) based on the number of samples for which data is available at each genomic position. This approach should work well as long as the amount of missing data is not excessive, but it will require some care in programming to accommodate site-wise variation in a_n efficiently. Another useful extension would be to allow for variation across loci in the global parameters ρ, η, and γ, say, by assuming locus-specific parameters are drawn from Beta (for ρ) or Gamma (for η and γ) distributions and estimating the hyper-parameters for these distributions from the data. This strategy should improve model fit considerably in cases of variable selection across loci, as with phylogenetic models that allow for rate variation among sites (Yang, 1994). A further extension would be to use a fully Bayesian approach and infer posterior distributions for the parameters of interest. This would also be fairly straightforward, but would most likely require Markov chain Monte Carlo sampling or variation Bayes approximations. These and other extensions would help further in using patterns of polymorphism and divergence to shed light on recent evolutionary processes, particularly in noncoding regions, and to improve predictions of the fitness effects of mutations across the genome.

Literature Cited

Andolfatto P. 2005. Adaptive evolution of non-coding DNA in Drosophila. Nature. 437:1149–1152.

Arbiza L, Gronau I, Aksoy BA, Hubisz MJ, Gulko B, Keinan A, Siepel A. 2012. Genome-wide inference of natural selection on human transcription factor binding sites. In prep. .

Bailey TL, Elkan C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In: Proc. 6th Int'l Conf. on Intelligent Systems for Molecular Biology. pp. 28–36.

Bierne N, Eyre-Walker A. 2004. The genomic rate of adaptive amino acid substitution in Drosophila. Mol. Biol. Evol. 21:1350–1360.

Boyko AR, Williamson SH, Indap AR, et al. (14 co-authors). 2008. Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet. 4:e1000083.

Bustamante CD, Fledel-Alon A, Williamson S, et al. (11 co-authors). 2005. Natural selection on protein-coding genes in the human genome. Nature. 437:1153–1157.

Bustamante CD, Nielsen R, Sawyer SA, Olsen KM, Purugganan MD, Hartl DL. 2002. The cost of inbreeding in Arabidopsis. Nature. 416:531–534.

Charlesworth J, Eyre-Walker A. 2008. The McDonald-Kreitman test and slightly deleterious mutations. Mol. Biol. Evol. 25:1007–1015.

Chernoff H. 1954. On the distribution of the likelihood ratio. Ann Math Statist. 25:573–578.

Chiaromonte F, Weber RJ, Roskin KM, Diekhans M, Kent WJ, Haussler D. 2003. The share of human genomic DNA under selection estimated from human-mouse genomic alignments. In: Cold Spring Harbor Symp Quant Biol. volume 68, pp. 245–254.

Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A. 2005. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. 15:901–913.

Dermitzakis ET, Clark AG. 2002. Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover. Mol Biol Evol. 19:1114–1121.

Eyre-Walker A, Keightley PD. 2009. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Mol. Biol. Evol. 26:2097–2108.

Eyre-Walker A, Woolfit M, Phelps T. 2006. The distribution of fitness effects of new deleterious amino acid mutations in humans. Genetics. 173:891–900.

Gerstein MB, Lu ZJ, Van Nostrand EL, et al. (131 co-authors). 2010. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. Science. 330:1775–1787.

Guigó R, Dermitzakis ET, Agarwal P, et al. (11 co-authors). 2003. Comparison of mouse and human genomes followed by experimental verification yields an estimated 1,019 additional genes. Proc Natl Acad Sci USA. 100:1140–1145.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 5:e1000695.

Hernandez RD. 2008. A flexible forward simulator for populations subject to selection and demography. Bioinformatics. 24:2786–2787.

Hubisz MJ, Pollard KS, Siepel A. 2011. PHAST and RPHAST: Phylogenetic analysis with space/time models. Briefings in Bioinformatics. 12:41–51.

Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature. 423:241–254.

Kondrashov AS, Crow JF. 1993. A molecular approach to estimating the human deleterious mutation rate. Hum. Mutat. 2:229–234.

Lehmann EEL, Casella G. 1998. Theory of point estimation. Springer.

Lunter G, Ponting CP, Hein J. 2006. Genome-wide identification of human functional DNA using a neutral indel model. PLoS Comput. Biol. 2:e5.

Mackay TF, Richards S, Stone EA, et al. (52 co-authors). 2012. The Drosophila melanogaster Genetic Reference Panel. Nature. 482:173–178.

Margulies EH, Blanchette M, NISC Comparative Sequencing Program, Haussler D, Green ED. 2003. Identification and characterization of multi-species conserved sequences. Genome Res. 13:2507–2518.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in Drosophila. Nature. 351:652–654.

Moses AM, Pollard DA, Nix DA, Iyer VN, Li XY, Biggin MD, Eisen MB. 2006. Large-scale turnover of functional transcription factor binding sites in Drosophila. PLoS Comput. Biol. 2:e130.

Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature. 420:520–562.

Myers RM, Stamatoyannopoulos J, Snyder M, et al. (11 co-authors). 2011. A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol. 9:e1001046.

Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 20:110–121.

Pollard KS, Salama SR, Lambert N, et al. (11 co-authors). 2006. An RNA gene expressed during cortical development evolved rapidly in humans. Nature. 443:167–172.

Prabhakar S, Visel A, Akiyama JA, et al. (13 co-authors). 2008. Human-specific gain of function in a developmental enhancer. Science. 321:1346–1350.

Roy S, Ernst J, Kharchenko PV, et al. (269 co-authors). 2010. Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science. 330:1787–1797.

Sawyer SA, Hartl DL. 1992. Population genetics of polymorphism and divergence. Genetics. 132:1161–1176.

Schmidt D, Wilson MD, Ballester B, et al. (13 co-authors). 2010. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science. 328:1036–1040.

Self S, Liang K. 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. J. Am. Stat. Assoc. 82:605–610.

Siepel A, Bejerano G, Pedersen JS, et al. (16 co-authors). 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 15:1034–1050.

Siepel A, Diekhans M, Brejova B, et al. (18 co-authors). 2007. Targeted discovery of novel human exons by comparative genomics. Genome Res. 17:1763–1773.

Smith NG, Eyre-Walker A. 2002. Adaptive protein evolution in Drosophila. Nature. 415:1022–1024.

Stoletzki N, Eyre-Walker A. 2011. Estimation of the neutrality index. Mol. Biol. Evol. 28:63–70.

The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. Nature. 467:1061–1073.

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. Theor Popul Biol. 7:256–276.

Williamson SH, Hernandez R, Fledel-Alon A, Zhu L, Nielsen R, Bustamante CD. 2005. Simultaneous inference of selection and population growth from patterns of variation in the human genome. Proc. Natl. Acad. Sci. U.S.A. 102:7882–7887.

Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol. 39:306–314.

Yi X, Liang Y, Huerta-Sanchez E, et al. (70 co-authors). 2010. Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude. Science. 329:75–78.

Tables

Table 1. Model parameters

Parameter	Type	Description
$\boldsymbol{\lambda}^{\mathbf{O}} = \{\lambda_b^O\}_{b \in B}$	neutral	Block-specific neutral scaling factor for the outgroup portion of the phylogeny, used when computing the prior distributions for the deep ancestral states, $\{P(Z_i \mid O_i, \lambda_b^O)\}_i$
$\boldsymbol{\lambda} = \{\lambda_b\}_{b \in B}$	neutral	Block-specific neutral scaling factor for divergence
$\boldsymbol{\theta} = \{\theta_b\}_{b \in B}$	neutral	Block-specific neutral polymorphism rate
$\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)$	neutral	Proportions of the three classes of derived allele frequency, $(0, f)$, $[f, 1-f]$, and $(1-f, 1)$, within neutral polymorphic sites
ρ	selection	Fraction of sites under selection within functional elements
η	selection	Ratio of divergence rate at selected sites to local neutral divergence rate
γ	selection	Ratio of polymorphism rate at selected sites to local neutral polymorphism rate

Table 2. Model variables associated with site i

Variable	Type	Description
O_i	observed	Column of the multiple genome alignment of outgroup species
X_i^{maj}	observed	Base for major allele in target population
X_i^{min}	observed	Base for minor allele in target population (NA for monomorphic sites)
Y_i	observed	MAF class for site i: 'M' for monomorphic sites (MAF=0) 'L' for polymorphic sites with MAF $< f$ 'H' for polymorphic sites with MAF $\geq f$
S_i	hidden	Selection class: 'neut' for neutral sites 'sel' for sites under selection
Z_i	hidden	Deep ancestral allele representing the root of the branch leading to the target population in the phylogeny spanning the outgroup species
A_i	hidden	Ancestral allele at the most recent common ancestor (MRCA) of the target population

Table 3. Conditional distribution table for $\boldsymbol{P(X_i \mid S_i, Z_i,\ \zeta)}$

s	y	$z,\ x_i^{maj},\ x_i^{min}$	$P\left(X_i = (x^{maj}, x^{min}, y) \mid S_i = s, Z_i = z,\ \boldsymbol{\zeta}\right)$
neut	M	$z = x^{maj},\ x^{min} = \text{NA}$	$(1-\lambda_b t)(1-\theta_b a_n)$
neut	M	$z \neq x^{maj},\ x^{min} = \text{NA}$	$\frac{1}{3}\lambda_b t(1-\theta_b a_n)$
neut	L	$z = x^{maj},\ x^{min} = *$	$\left((1-\lambda_b t)\beta_1 + \frac{1}{3}\lambda_b t\beta_3\right)\frac{1}{3}\theta_b a_n$
neut	L	$z = x^{min},\ x^{maj} = *$	$\left((1-\lambda_b t)\beta_3 + \frac{1}{3}\lambda_b t\beta_1\right)\frac{1}{3}\theta_b a_n$
neut	L	$z \notin \{x^{maj}, x^{min}\}$	$\frac{1}{3}\lambda_b t\left(\beta_1 + \beta_3\right)\frac{1}{3}\theta_b a_n$
neut	H	$z \in \{x^{maj}, x^{min}\}$	$\left(1-\lambda_b t + \frac{1}{3}\lambda_b t\right)\beta_2\frac{1}{3}\theta_b a_n$
neut	H	$z \notin \{x^{maj}, x^{min}\}$	$\frac{2}{3}\lambda_b t\beta_2\frac{1}{3}\theta_b a_n$
sel	M	$z = x^{maj},\ x^{min} = \text{NA}$	$(1-\eta\lambda_b t)(1-\gamma\theta_b a_n)$
sel	M	$z \neq x^{maj},\ x^{min} = \text{NA}$	$\frac{1}{3}\eta\lambda_b t$
sel	L	$z = x^{maj},\ x^{min} = *$	$(1-\eta\lambda_b t)\frac{1}{3}\gamma\theta_b a_n$
sel	L	$z \neq x^{maj},\ x^{min} = *$	0
sel	H	$x^{maj} = *,\ x^{min} = *,\ z = *$	0

Table 4. Selection in human transcription factor binding sites (TFBS)

TF	BS length	element sites	flanking sites	ρ	$\mathbb{E}[D_\text{p}]$	$\mathbb{E}[P_\text{w}]$
BRCA1	15	142,450	73,656,594	0.43 ± 0.04 ***	0.00 ± 0.19	0.97 ± 0.21 **
CTCF	13	802,273	408,024,473	0.34 ± 0.02 ***	0.00 ± 0.10	1.21 ± 0.11 **
GATA2	7	229,781	200,350,965	0.32 ± 0.05 ***	1.11 ± 0.24 **	0.86 ± 0.23 **
SUZ12	11	137,174	183,986,075	0.32 ± 0.09 *	0.86 ± 0.43	1.44 ± 0.43 *

[a] Number of bases in the motif computed by MEME for the TF.

[b] Total Number of sites analyzed within binding sites of the TF, after filtering.

[c] Total number of non-filtered putative neutral sites within 10 kb flanking regions of binding sites.

[d] Estimates of ρ with curvature-based standard errors.

[e] Posterior expected values of D_p normalized per 1,000 bp, with curvature-based standard errors.

[f] Posterior expected values of P_w normalized per 1,000 bp, with curvature-based standard errors.

[*] Estimates found to be significantly greater than zero, (* $p = 0.01$; ** $p = 0.001$; *** $p = 10^{-5}$).
P-values estimated using χ^2 with three degrees of freedom for testing $\rho > 0$ and χ^2 with one degree of freedom for testing $\eta > 0$ and $\gamma > 0$.

Figure Legends

Figure 1. Schematic description of the P&D model. (A) The P&D model consists of three components: a phylogenetic model (gray) that relates the outgroup sequence data, $\mathbf{O}$, with the deep ancestral genome ,Z, a divergence model (blue) that relates Z with the sequence, A, representing the most recent common ancestor (MRCA) of the population sample at any given site, and a population genetic model for polymorphism (red) that relates A with the population sequence data, $\mathbf{X}$. (B) The data consists of a collection of sites in functional elements (set E; gold) and putative neutral sites in flanking regions of these elements (set F; dark gray). Variation in rates of polymorphism and divergence along the genome is considered by grouping nearby elements and their flanks into a genomic blocks and allowing different neutral polymorphism and divergence rates at different blocks. Some sites within each block are masked (light gray). Sequence data consists of a collection of individual genomes sampled from the target population and several genomes of outgroup species closely related to the target population. All genomes are aligned to a reference genome of the target population, and single nucleotide differences from that reference are indicated by red ticks. The outgroup genomes are used to probabilistically infer the ancestral genome, Z, and the MRCA sequence, A, which are then used to determine divergent sites and polarize polymorphic sites. Likely inference of Z and A are depicted in the figure, and examples of six sites are given (numbered dotted vertical lines). Sites 3 and 4 are inferred to be divergent, and sites 1,2, and 6 are low-frequency polymorphic sites, and site 5 is a high-frequency polymorphic site.

Figure 2. Graphical model for P&D at site $\boldsymbol{i}$. The three components of the model are highlighted in color, as in Fig. 1A: phylogenetic (gray), divergence (blue), and population genetic (red). Observed variables are represented by solid circles and hidden variables are represented by empty circles. Variables $X_i = (X_i^{maj}, X_i^{min}, Y_i)$ and O_i represent the sequence data in the target population and outgroup sequences, respectively. Variable S_i represents the selection class, and variables Z_i and A_i represent two unknown ancestral states. Conditional dependence between the variables is indicated by directed edges, and model parameters are portrayed next to the edges that represent the associated conditional distributions. The selection parameters $\boldsymbol{\zeta}_{\text{sel}} = (\rho, \eta, \gamma)$ are highlighted in green.

Figure 3. Results on Simulated Data. (A) Parameter estimates for four different collections of 20,000 simulated elements with different proportions for strong positive (SP), strong negative (SN), and weak negative (WN) selection. The proportion for each of these is indicated at the bottom together with the total proportion for all selected sites. For each collection the true values of ρ, D_{p}, and P_{w} are indicated (solid bars) alongside the estimates obtained by INSIGHT, which are shown with confidence intervals representing curvature-based standard errors (see **Methods**). The P&D counts, D_{p} and P_{w}, are normalized per 1,000 base pairs (kbp). For comparison, we are also showing the polymorphism-based estimates for ρ ($\hat{\rho}_{\mathrm{Poly}}$; '+' labels; see **Methods**) and the divergence-based estimates ($\hat{\rho}_{\mathrm{Div}}$; solid squares), as well as the estimates of D_{p} based on the McDonald-Kreitman framework ($\hat{D}_{\text{p-MK}}$; '×' labels). (B) Seven collections of 10,000 elements were simulated each under four different demographic scenarios with varying degrees of complexity (Supplementary Table S1). The relative estimation error for ρ, D_{p}, and P_{w} is measured by the difference between the estimates and true values normalized by the true value. Box plots represent the distribution of relative error across the seven collections for each demographic scenario. The relative estimation error of the simple site-count-based estimates, $\hat{\rho}_{\mathrm{Poly}}$, $\hat{\rho}_{\mathrm{Div}}$, and $\hat{D}_{\text{p-MK}}$ are shown for comparison. (C) Selection inference was performed on each of the seven collections used in Fig.3B assuming different thresholds between low- and high-frequency polymorphisms. The relative estimation errors (left) and curvature-based standard errors (right) are plotted against the frequency threshold used in inference. Each boxplot describes the distribution across the seven collections for a given frequency threshold.

Figure 4. Analysis of human transcription factor binding sites (TFBSs). (A) Control study using a set of 500 collections of "neutral" elements extracted from putative neutral regions. For each of the 500 collections, we obtained an INSIGHT estimate of ρ, and a test statistic for the hypothesis that $\rho > 0$. The distribution of estimated ρ values (left) is showed next to the distribution of the test statistics (right). The χ^2 distribution with three degrees of freedom is portrayed (red) together with two cutoff points for p-values of 0.01 and 0.05. Eight out of the 500 collections (1.6%) have test statistics that exceed the $p = 0.01$ cutoff, and 23 of the 500 collections (4.6%) have test statistics that exceed the $p = 0.05$ cutoff. (B) Estimates of ρ for CTCF (left) and SUZ12 (right) obtained by fitting the model using different frequency thresholds to distinguish low- and high-frequency polymorphisms. Horizontal lines representing the simple estimates of ρ based on counts of polymorphic sites ($\hat{\rho}_{\text{Poly}}$; red) and counts of divergent sites ($\hat{\rho}_{\text{Div}}$; blue) are shown for comparison. Estimate based on the number of high-frequency polymorphisms (dashed red) are also plotted as a function of the frequency cutoff (see text). (C) The motif inferred by MEME for GATA2 is shown above estimates of ρ (left axis), D_{p}, and P_{w} (right axis), obtained separately for each position along the motif. Confidence intervals represent curvature-based standard errors, and D_{p}, and P_{w} are normalized per kbp. Positions found to have ρ, $\mathbb{E}[D_{\text{p}}]$, and $\mathbb{E}[P_{\text{w}}]$ significantly greater than zero ($p = 0.01$) are indicated by labels 'S', 'p', and 'w', respectively (bottom line). Significance is assesses using a χ^2 distribution with three degrees of freedom for ρ and a χ^2 distribution with a single degree of freedom for $\mathbb{E}[D_{\text{p}}]$ and $\mathbb{E}[P_{\text{w}}]$. The portion of the motif with IC$> \frac{1}{2}$ is highlighted (gray). Five of the seven positions in this trimmed version of the motif have significant evidence for selection. Positions 6 and 7 show significant evidence of positive selection, and positions 6 and 8 show significant evidence of weak negative selection.

Figures

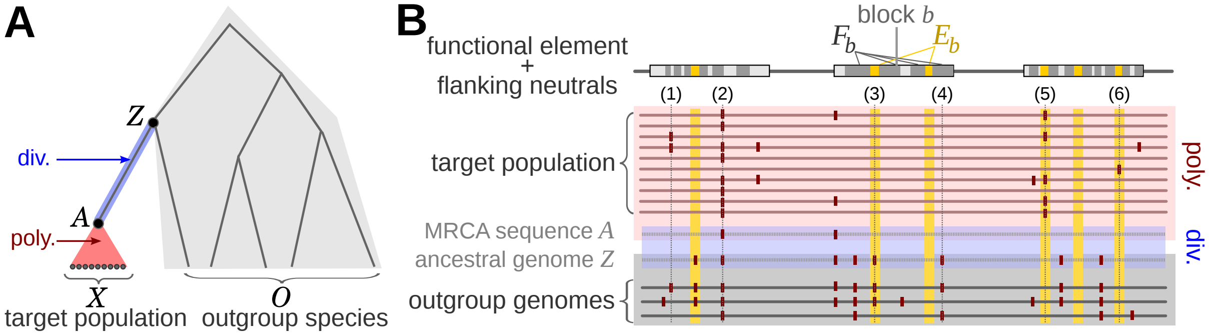

Figure 1

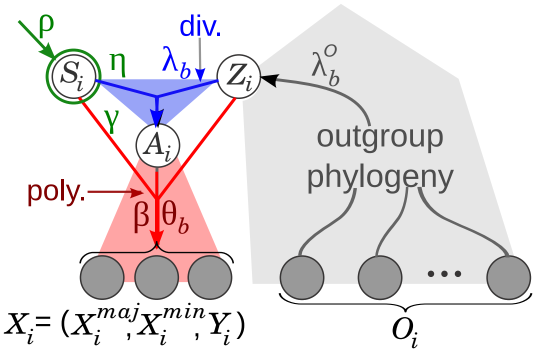

Figure 2

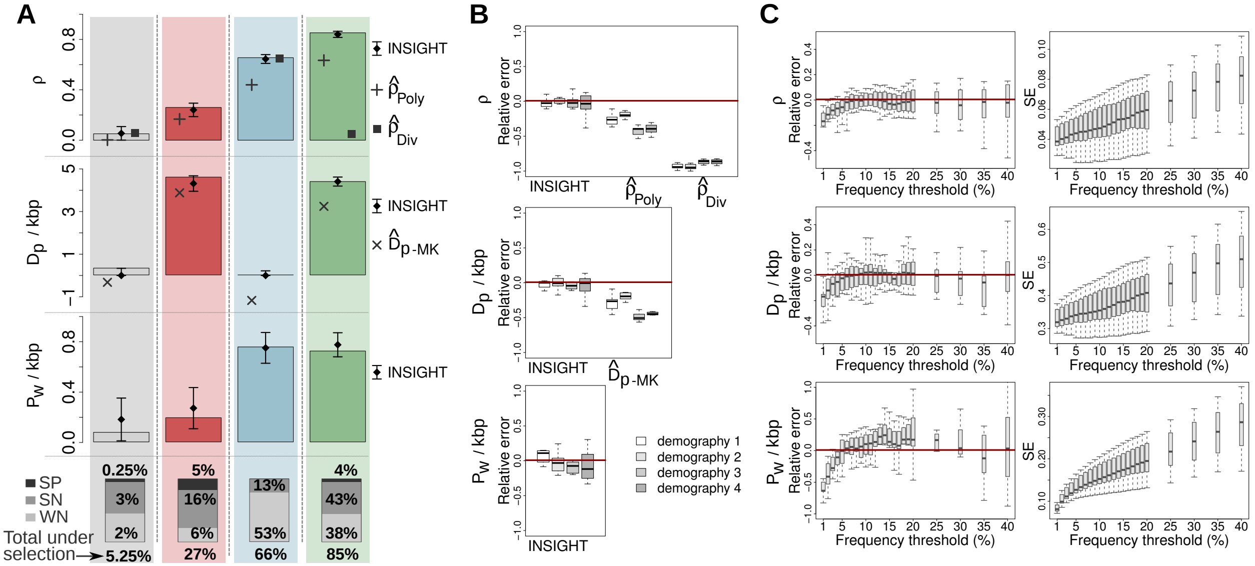

Figure 3

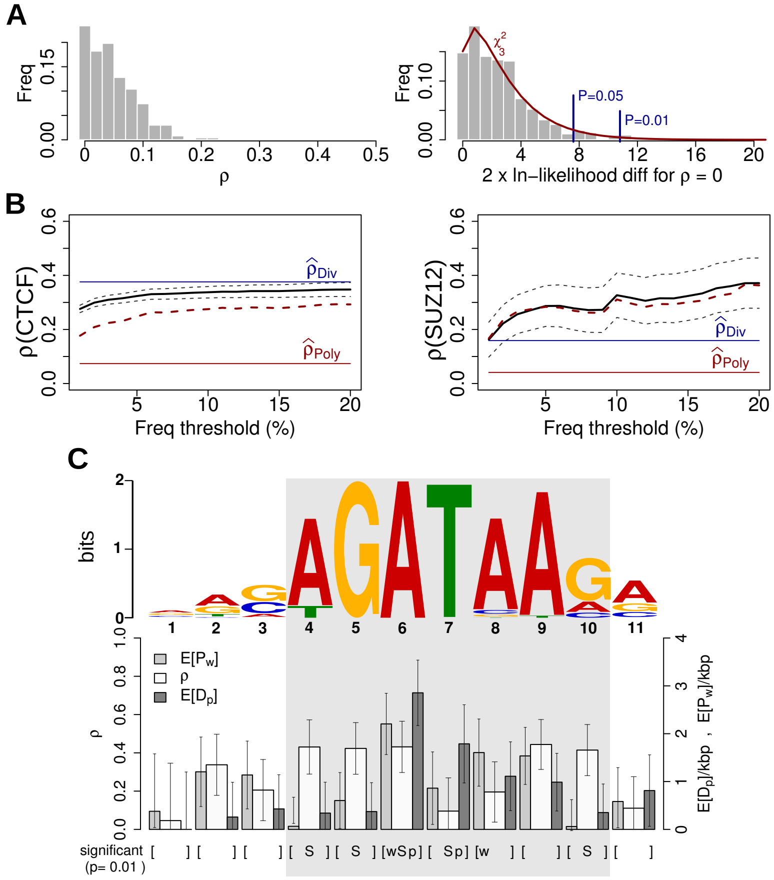

Figure 4

Supplementary Tables

Table S1. Demographic scenarios

Time [a]	scenario 1 [b]	scenario 2 [c]	scenario 3 [d]	scenario 4 [d]
220 kya [e]	–	1.23x [h]	1.23x [h]	1.23x [h]
140 kya [f]	–	–	0.17x [h]	0.17x [h]
20.8 kya [g]	–	–	0.476x [h]	0.242x [h]
20.8 kya [g]	–	–	exp(79.8) [i]	exp(109.7) [i]

[a] Time of change in N_e (kya = 1,000 years ago). All demographic scenarios start with an ancestral N_e of 10,000 after divergence from chimpanzee, 6.5 mya. Demographic scenarios follow the model suggested by Gutenkunst et al. (2009).

[b] Scenario 1 no demographic changes throughout history.

[c] Scenario 2 corresponds to the demographic history of an African population with a single moderate population expansion.

[d] Scenario 3 & 4 correspond to the demographic histories of a European and East Asian population, resp., each with a moderate population expansion followed by two population bottlenecks and an exponential expansion.

[e] Moderate population expansion in the African population ancestral to all current human populations.

[f] Divergence point of an ancestral Eurasian population from an ancestral African population associated with a strict population bottleneck in the ancestral Eurasian population.

[g] Divergence of European and East Asian population associated with additional bottlenecks in both ancestral populations followed by exponential expansion.

[h] Instantaneous population size increase or decrease by a given multiplicative factor.

[i] Exponential population size expansion at a given rate expressed as $\log(N_e^{\text{final}}/N_e^{\text{initial}})/\text{time}$, where time is in units of $2N_e$ generations.